UNIVERSITÉ DE MONTRÉAL

CENTRE DE RECHERCHE SUR LES TRANSPORTS

CASE POSTALE 6128, SUCCURSALE A, MONTRÉAL, QUÉBEC, H3C 3J7

SOME COMMENTS ON WOLFE'S "AWAY STEP"

by

Jacques Guélat[1] and Patrice Marcotte[2]

[1] Centre de recherche sur les transports, Université de Montréal.

[2] Collège Militaire Royal de Saint-Jean et Centre de recherche sur les transports, Université de Montréal.

Cet article a été publié grâce à une subvention du Fonds F.C.A.C. pour l'aide et le soutien à la recherche et du C.R.S.N.G. (Canada).

Université de Montréal

Centre de recherche sur les transports - Publication #369

Septembre 1984 - Révisé en octobre 1985

ABSTRACT

We give a detailed proof, under slightly weaker conditions on the objective function, that a modified Frank-Wolfe algorithm based on Wolfe's "away step" strategy, can achieve geometric convergence, provided a strict complementarity assumption holds.

Key words : convex programming, Frank-Wolfe algorithm.

RÉSUMÉ

Cet article vise à démontrer rigoureusement qu'un algorithme de FW modifié et inspiré de la variante "away step" proposée par Wolfe possède, sous l'hypothèse de complémentarité stricte, un taux de convergence géométrique.

Mots clés : programmation convexe, algorithme de Frank-Wolfe.

1. INTRODUCTION

Consider the mathematical programming problem :

$$\min_{x \in S} f(x) \qquad (P)$$

where f is a convex, continuously differentiable function, and $S = \{x \in \mathbb{R}^n \mid Ax=b,\ x \geq 0\}$ a nonempty bounded polyhedron. Throughout the paper we will denote by R the set of extreme points of S. In [2] Frank and Wolfe proposed the following algorithm for solving P :

Algorithm FW

Let $x^1 \in S$.

$k \leftarrow 1$.

1. Solving a linearized problem

Let $T(x^k) \triangleq \arg\min_{z \in S} f(x^k) + (z-x^k)^T \nabla f(x^k) = \arg\min_{z \in S} z^T \nabla f(x^k)$.

Let $y \in R \cap T(x^k)$, $d = y - x^k$.

2. Line search

Let $\gamma \in \arg\min_{t \in [0,1]} f(x^k + td)$.

3. Update

$x^{k+1} \leftarrow x^k + \gamma d$.

$k \leftarrow k+1$ and return to 1.

It is straightforward to show that if $\gamma = 0$ at step 2, then $x^k \in T(x^k)$ and x^k is optimal for P. In practice, the algorithm has to be stopped after some convergence criterion is satisfied. From the convexity of f, we have :

$$f(x^*) \geq f(x^k) + (y-x^k)^T \nabla f(x^k).$$

Therefore, the term $(x^k-y)^T\nabla f(x^k)$, where y is solution to the linearized problem, provides an upper bound on the difference between the objective evaluated, respectively, at the current iterate and at the (unknown) optimum. This quantity is sometimes referred to as the gap associated with problem P at the current iterate, and can be conveniently utilized as a measure of proximity to the optimum. For instance, in many applications, f is strictly positive at the optimum, and the algorithm is stopped as soon as the inequality $(x^k-y)^T\nabla f(x^k) \leq \varepsilon f(x^k)$ is satisfied, for some small positive constant ε.

Global convergence of the algorithm can be proved by showing that the algorithm map is closed, (see Zangwill [6] or Luenberger [3]) or by bounding from below the decrease of the objective at each iteration (see next section).

In the FW algorithm, descent directions are always directed toward extreme points of S. When one gets close to the optimum, and when this optimum is a point on the boundary of S, these directions become more and more orthogonal to the gradient vector, without reaching the optimal face[1], resulting in a poor (sublinear) convergence rate. To remedy this situation, Wolfe [5] suggested enlarging the set of admissible directions, by including directions pointing "away" from extreme points. Wolfe sketched a proof that the modified algorithm identifies the set of active constraints (optimal face) in a finite number of iterations, and achieves geometric convergence, provided the objective function is twice continuously differentiable, strongly convex and that strict complementarity holds at x*. In this paper we will give complete proofs of Wolfe's main results, under slightly weaker conditions.

---

[1] A face of S is a convex subset C such that every closed line segment in S with a relative interior point in C lies entirely in C. See Rockafellar [4] for details.

2. BASIC CONVERGENCE RESULTS

Throughout the paper, we will make use of the following three assumptions.

Assumption 1. ∇f is Lipschitz-continuous on S with Lipschitz constant λ_2, i.e. : $\|\nabla f(y) - \nabla f(x)\| \leq \lambda_2 \|y-x\|$ for all x,y in S. (1)

Assumption 2 (strong convexity)[2]. There exists a positive constant λ_1 such that : $(y-x)^T(\nabla f(y)-\nabla f(x)) \geq \lambda_1 \|y-x\|^2$. (2)

Assumption 3 (strict complementarity)

Let x* be optimal for P and T* be the smallest face containing x*. Then : $(y-x^*)^T\nabla f(x^*) = 0 \iff y \in T^*$.

Under assumption 2, the optimal solution x* is unique.

The next theorem provides a lower bound on the rate of convergence of the FW algorithm.

Theorem 1. Let x* be optimal for P and $\{x^k\}_k$ be a sequence generated by FW. If assumption 1 holds, then there exists an index K and constants L and ξ such that :

$$f(x^k) - f(x^*) \leq \frac{L}{k+\xi} \tag{3}$$

for $k \geq K$.

Proof. By convexity of f and definition of d, one has :

$$f(x^k) - f(x^*) \leq -d^T\nabla f(x^k) \tag{4}$$

Let $t \in [0,1]$. By the mean value theorem :

---

[2] See Auslender [1] for characterizations of various convexity concepts when f is differentiable.

$$f(x^k+td)-f(x^k) = td^T\nabla f(x^k+t'td) \quad \text{with } t'\in[0,1]$$

$$= td^T\nabla f(x^k) + td^T(\nabla f(x^k+t'td) - \nabla f(x^k))$$

$$\leq \frac{1}{2}\, td^T\nabla f(x^k) + \frac{t}{2}(f(x^*)-f(x^k)) + \lambda_2 t^2\|d\|^2$$

by (1) and (4)

$$\leq \frac{1}{2}\, td^T\nabla f(x^k) + \frac{t}{2}(f(x^*)-f(x^k) + 2t\lambda_2 D^2)$$

where D is the diameter of S.

$$\leq \frac{1}{2} td^T\nabla f(x^k)$$

if $t \leq m_k \overset{\Delta}{=} \text{Min}\left\{1, \frac{f(x^k)-f(x^*)}{2\lambda_2 D^2}\right\}$.

Thus : $$f(x^{k+1})-f(x^*) \leq \left(1 - \frac{m_k}{2}\right)(f(x^k)-f(x^*)). \qquad (5)$$

Since the sequences $\{f(x^k)\}_k$ and $\{m_k\}_k$ are decreasing and bounded, they admit limits f^∞ and m^∞, respectively. By taking limits on both sides of (5), one finds :

$$f^\infty - f(x^*) \leq \left(1 - \frac{m^\infty}{2}\right)(f^\infty - f(x^*))$$

This implies that either $f^\infty = f(x^*)$ or $m^\infty = 0$, which is equivalent. In particular global convergence is proved.

Now let K be the smallest index such that $\frac{f(x^K)-f(x^*)}{2\lambda_2 D^2} \leq 1$.

Then, from (5) : $\frac{m_{k+1}}{2} \leq (1 - \frac{m_k}{2}) \frac{m_k}{2}$ for $k > K$

and $\frac{2}{m_{k+1}} \geq 1 + \frac{2}{m_k}$

or : $\frac{2}{m_k} \geq k - K + \frac{2}{m_K}$.

Thus : $m_k \leq \frac{2}{k-K+2/m_K} = \frac{2}{k+\xi}$ with $\xi = \frac{2}{m_K} - K$.

Setting $L = 4\lambda_2 D^2$, we get the desired result. Q.E.D.

In a particular instance, the convergence rate can be shown to be linear.

Theorem 2. Suppose assumptions 1 and 2 are satisfied and let the optimal solution x^* be in the relative interior of S. Then the sequence $\{x^k\}_k$ generated by FW converges geometrically to x^*.

Proof. If the relative interior of S is empty, there is nothing to prove. Otherwise, let H_S be the smallest affine variety containing S.

Since x^* (unique by assumption 2) is in the relative interior of S, there must exist some open ball $B(x^*,2\varepsilon)$ around x^* such that $B(x^*,2\varepsilon) \cap H_S \subset S$, and an index K such that $x^k \in B(x^*,\varepsilon)$ for all $k > K$. From that iteration on, the problem is equivalent to an unconstrained optimization problem relative to H_S. Therefore, in the rest of the proof, we use ∇f instead of the restriction of ∇f to H_S.

Now let $k>K$, $d=y-x^k$ be a FW direction at x^k ($x^k \neq x^*$) and $t \in [0,1]$. From assumptions 1 and 2 :

$$\lambda_1 \|x^k+td-x^k\|^2 \leq (x^k+td-x^k)^T(\nabla f(x^k+td)-\nabla f(x^k)) \leq \lambda_2 \|x^k+td-x^k\|^2. \tag{6}$$

Let t^* be the unique value of t for which the minimum value of f is achieved, i.e. :

$$d^T \nabla f(xk+t^*d) = 0.$$

We have that :

$$x^{k+1} = x^k + t^* d \in B(x^*, \varepsilon)$$

by definition of K.

Dividing both sides of (6) by t and integrating from 0 to t we obtain :

$$\lambda_1 \frac{t^2}{2} \|d\|^2 + t d^T \nabla f(x^k) \leq f(x^k+td) - f(x^k) \leq \lambda_2 \frac{t^2}{2} \|d\|^2 + t d^T \nabla f(x^k). \tag{7}$$

Since $x^{k+1} = x^k + t^* d$ achieves the minimum of $f(x^k+td)$ for $t > 0$, $f(x^{k+1})-f(x^k)$ must be bounded by the respective minima of the quadratic terms of (7), i.e. :

$$\frac{(d^T \nabla f(x^k))^2}{2\lambda_2 \|d\|^2} \leq f(x^k)-f(x^{k+1}) \leq \frac{(d^T \nabla f(x^k))^2}{2\lambda_1 \|d\|^2}.$$

Or

$$\frac{\|\nabla f(x^k)\|^2 \cos^2\theta}{2\lambda_2} \leq f(x^k)-f(x^{k+1}) \leq \frac{\|\nabla f(x^k)\|^2 \cos^2\theta}{2\lambda_1} \tag{8}$$

where θ is the angle between the vectors d and $\nabla f(x^k)$. Relation (8) is valid for any descent direction d; in particular, taking $d = \nabla f(x^k)$ and $d = x^*-x^k$ we find :

$$\frac{\|\nabla f(x^k)\|^2}{2\lambda_2} \leq f(x^k)-f(x^*) \leq \frac{\|\nabla f(x^k)\|^2}{2\lambda_1}. \tag{9}$$

Now divide the three terms of (8) by the (reversely) respective terms of (9), to find :

$$\frac{\lambda_1}{\lambda_2} \cos^2\theta \leq \frac{f(x^k)-f(x^{k+1})}{f(x^k)-f(x^*)} \leq \frac{\lambda_2}{\lambda_1} \cos^2\theta . \tag{10}$$

It remains to check that, for any FW direction d, $\cos^2\theta$ is bounded away from zero.

We have that $x^k - \varepsilon \frac{\nabla f(x^k)}{\|\nabla f(x^k)\|}$ is in S since $B(x^*,2\varepsilon) \cap H_S$ lies in S, and x^k is in $B(x^*,\varepsilon) \cap H_S$. Let $d=y-x^k$ be a Frank-Wolfe direction.

Then : $(y-x^k)^T \nabla f(x^k) \leq -\varepsilon \|\nabla f(x^k)\|$.

Therefore : $|\cos\theta| = \frac{|(y-x^k)^T \nabla f(x^k)|}{\|y-x^k\| \, \|\nabla f(x^k)\|} \geq \frac{\varepsilon}{D} > 0$. Q.E.D.

The next result shows that a FW sequence cannot in general be expected to converge geometrically to a solution.

Theorem 3. Suppose assumptions 1 and 2 hold and that :

- the (unique) solution x* lies on the boundary of S.
- $x^* \notin R$ (x^* is not an extreme point of S).
- $x^K \notin T^*$ (smallest face containing x^*) for some index K.

Then for any positive constant δ, the relation $f(x^k)-f(x^*) > \frac{1}{k^{1+\delta}}$ holds for infinitely many indices k.

Proof. See Wolfe [5].

3. A MODIFIED FRANK-WOLFE ALGORITHM

Algorithm MFW.

Let $x^1 \in S$.

$k \leftarrow 1$.

1. Solving first linearized problem (Toward Step)

Let $y_T \in T(x^k) \cap R$ and $d_T = y_T - x^k$. (LPT)

2. Solving the second linearized problem (Away Step)

Let $A(x^k) \triangleq \arg\max_z f(x^k) + (z-x^k)^T \nabla f(x^k)$ (LPA)

s.t. $z \in S$ and $z_i = 0$ if $x_i^k = 0$.

Set : $y_A \in A(x^k) \cap R$ ($y_A = x^k$ if $A(x^k) \cap R = \emptyset$).

$d_A = x^k - y_A$.

$\alpha_A = \max_{\beta \geq 0} \{\beta \mid x^k + \beta d_A \in S\} > 0$.

3. Choosing a descent direction

If $d_T^T \nabla f(x^k) < d_A^T \nabla f(x^k)$ then $d = d_T$

else $d = d_A$

(if equality holds, d_A can be chosen as well).

4. Line search

Let $\gamma \in \arg\min_{t \in [0,\alpha]} f(x^k + td)$

where $\alpha = 1$ if $d = d_T$ and $\alpha = \alpha_A$ otherwise.

5. Update

$x^{k+1} \leftarrow x^k + \gamma d$.

$k \leftarrow k + 1$ and return to 1.

Remark 1 : In Wolfe's paper, the additionnal constraint $\{z_i=0 \text{ if } x_i^k=0\}$ of LPA is not included, so that stepsizes of zero length can occur at non-optimal points.

Remark 2 : At step 3 of algorithm MFW, Wolfe chooses d_A whenever $|y_A^T \nabla f(x^k)| > |y_T^T \nabla f(x^k)|$. This criterion is inadequate, since it can lead to nondescent directions. Indeed, consider the quadratic programming problem :

$$\underset{x \in S}{\text{Min}} \frac{1}{2}(x_1^2+x_2^2)$$

where S is the trapezoid illustrated in figure 1.

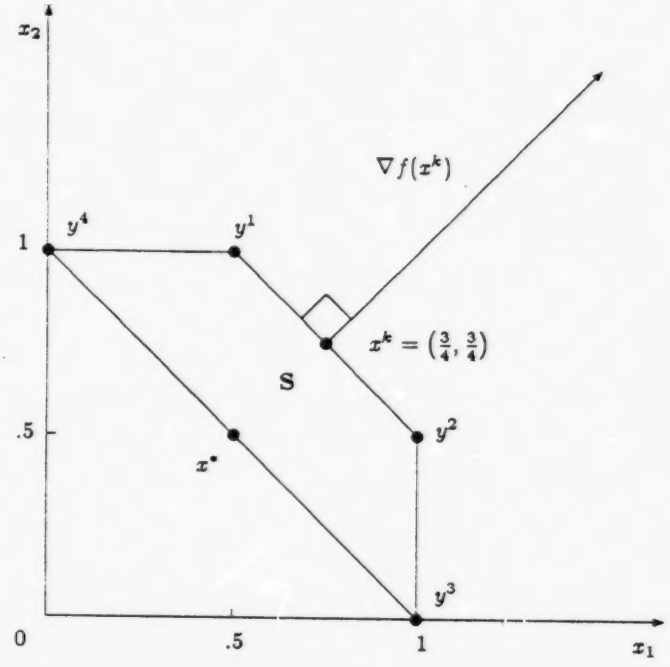
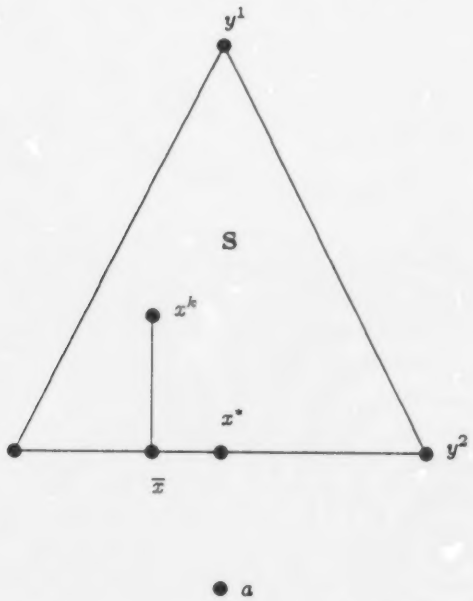


Figure 1

For $x^k = (\frac{3}{4},\frac{3}{4})^T$, one has : $y_T = y^3$ or y^4 and $y_A = y^1$ or y^2 with $|y_T^T \nabla f(x^k)| = \frac{3}{4}$ and $|y_A^T \nabla f(x^k)| = \frac{9}{8} > \frac{3}{4}$. However : $d_A^T \nabla f(x^k) = 0$; thus d_A is not a descent direction. This situation is not pathological : if the initial point x^1 lies on the segment $[y^1,y^2]$, then, according to Wolfe's rule, away steps will always be performed, and the sequence of iterates will converge to the nonoptimal point $(\frac{3}{4},\frac{3}{4})^T$.

It is easy to show that the algorithmic map corresponding to MFW is not closed (see figure 2). To prove global convergence, we will show that infinitely many steps will occur, each one resulting in an objective function decrease of the form given by expression (5), thus ensuring global convergence.



$f(x) = \|x-a\|^2$

$x^k \to \bar{x}$

$d(x^k) = x^k - y^1 \to \bar{x} - y^1$

but $d(\bar{x}) = y^2 - \bar{x}$

Figure 2 - Non closedness of the algorithmic map for MFW

Theorem 4. Let $\{x^k\}_k$ be a sequence generated by MFW. If assumption 1 holds, then : $\lim_{k\to\infty} f(x^k) = f(x^*)$, the optimal value of P.

Remark : In Wolfe [5], global convergence is assumed but not proved.

Proof. Suppose the algorithm is not finitely convergent, i.e. : $x^{k_1} \neq x^{k_2}$ whenever $k_1 \neq k_2$. At least one of the following situations must occur:

1. Infinitely many toward steps are performed and convergence follows immediately from (5).
2. Infinitely many away steps with step sizes less than the maximal step-size α_A are performed.

If this is the case, let $\{x^k\}_{k\in I}$ be a convergent subsequence and $\bar{x}$ its limit point. Suppose $f(\bar{x}) > f(x^*)$. Then there exists y in R and a positive constant C such that : $(y-\bar{x})^T\nabla f(\bar{x}) = -2C < 0$.

By continuity of ∇f, there exists a positive δ such that $\|x-\bar{x}\| < \delta$ implies : $(y-x)^T\nabla f(x) \leq -C < 0$.

Now let K be an index in I such that

$$\|x^k-\bar{x}\| < \delta \quad \text{for any } k>K,\ k\in I.$$

If, at x^k, an away step with stepsize less than α_A is performed, denote by ℓ the subsequent index in I.

Then : $f(x^k)-f(x^\ell) \geq f(x^k)-f(x^{k+1})$.

Since the stepsize is not maximal, relation (8) must hold, and :

$$f(x^k)-f(x^\ell) \geq \frac{1}{2\lambda_2}\|\nabla f(x^k)\|^2 \cos^2\theta$$

$$\geq \frac{1}{2\lambda_2}\left[\frac{(x^k-y_A)^T\nabla f(x^k)}{\|x^k-y_A\|}\right]^2$$

$$\geq \frac{C^2}{2\lambda_2 D^2}$$

$$> 0$$

in contradiction with the assumption $f(x^k) \underset{\substack{k\to\infty \\ k\in I}}{\to} f(\bar{x})$.

3. Neither 1 or 2 are satisfied.

There exists an index K such that $d=d_A$ for all $k>K$, and $\gamma=\alpha_A$ for those directions. This implies that the dimensionality of the minimal face containing x^k strictly decreases at each iteration, which is clearly impossible.

Q.E.D.

4. CONVERGENCE RATE OF MFW

Theorem 5. Under assumptions 1, 2 and 3, MFW identifies the set of active constraints in a finite number of iterations, and convergence towards the optimum x^* is geometric.

Proof. Let T^* denote the optimal face. We first show the existence of an index K such that for any $k > K$, the MFW direction at x^k is an away direction, unless x^k is already in T^*.

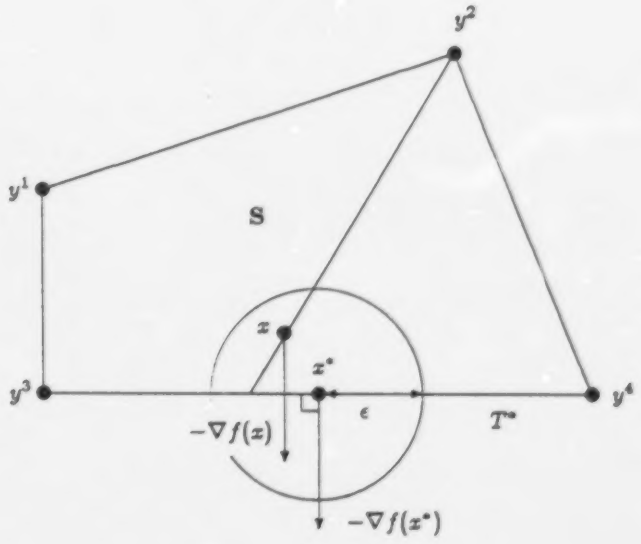


Figure 3

Let ε_j,c be positive constants such that

$$(y^j-x)^T\nabla f(x) > -\frac{c}{2} \quad \text{whenever } \|x-x^*\| < \varepsilon_j \text{ , } y^j \varepsilon R \cap T^*$$

$$(y^j-x)^T\nabla f(x) > c \quad \text{whenever } \|x-x^*\| < \varepsilon_j \text{ , } y^j \varepsilon R - T^*.$$

Let $\varepsilon = \underset{\{j|y^j \varepsilon R\}}{\text{Min}} \{\varepsilon_j\}$ and K an index such that $x^k \varepsilon B(x^*,\varepsilon)$ for all $k > K$. [3]

---

[3] The existence of such an index K is a direct consequence of the global convergence result of theorem 4 and of assumption 2.

If x^K is not in T* we have :

$$d_A^T \nabla f(x^K) \leq -c < d_T^T \nabla f(x^K), \quad \text{and } d = d_A.$$

We have : $x^{K+1} = x^K + \alpha d.$

Suppose that $\alpha < \alpha_A$. Then :

$$\begin{aligned} 0 = d^T \nabla f(x^{K+1}) &= (x^K - y^j)^T \nabla f(x^{K+1}) \\ &= (x^{K+1} - y^j)^T \nabla f(x^{K+1}) + (x^K - x^{K+1})^T \nabla f(x^{K+1}) \\ &= (x^{K+1} - y^j)^T \nabla f(x^{K+1}) - \alpha d^T \nabla f(x^{K+1}) \\ &\leq -c \quad \text{since} \quad x^{K+1} \in B(x^*, \varepsilon) \\ &< 0 . \end{aligned}$$

Thus we must have $\alpha = \alpha_A$ and the dimension of the face containing x^{K+1} is strictly smaller than the dimension of the face containing x^K. Starting at iteration K, the optimal face T* will therefore be reached in a finite number of iterations. This completes the proof of the first statement.

Now, once T* is reached, the iterates do not leave T* and the algorithm behaves as an unconstrained optimization algorithm on the affine variety described by the active constraints.

In the convergence proof of theorem 2, it is only required that $\cos\theta$ be negative and bounded away from zero. Since the away step strategy satisfies this requirement, the proof can be repeated almost word for word, after substituting for the gradient of f its projection onto the smallest affine variety containing T*.

It is worth noting that the constants λ_1 and λ_2 can be replaced by (maybe) smaller constants relative to the projection subspace.

Q.E.D.

5. CONCLUSION

In this paper, we have given detailed and complete proofs of convergence results about a modified version of the FW algorithm, slightly weakening the differentiability hypothesis assumed in Wolfe [5].

REFERENCES

[1] AUSLENDER, A. Optimisation - Méthodes numériques. Masson, Paris (1976).

[2] FRANK, M. and WOLFE, P. "An algorithm for quadratic programming". Naval Res. Log. Quart 3 (1956).

[3] LUENBERGER, D.G. Introduction to Linear and Nonlinear Programming. Addison Wesley, Reading, Mass. (1973).

[4] ROCKAFELLAR, R.T. Convex Analysis. Princeton University Press, Princeton, N.J. (1970).

[5] WOLFE, P. "Convergence theory in nonlinear programming", in Integer and Nonlinear Programming, Abadie, J.A. ed., North-Holland, Amsterdam,

[6] The Netherlands (1970).

ZANGWILL, W.I. Nonlinear Programming. A Unified Approach. Prentice-Hall, Englewood Cliffs, N.J. (1969).